DIALOGUE ACT RECOGNITION APPROACHES

Pavel Král

Dept. of Computer Science & Engineering
University of West Bohemia
Plzeň, Czech Republic
e-mail: pkral@kiv.zcu.cz

Christophe Cerisara

LORIA UMR 7503
BP 239 - 54506 Vandoeuvre
France
e-mail: cerisara@loria.fr

Abstract. This paper deals with automatic dialogue act (DA) recognition. Dialogue acts are sentence-level units that represent states of a dialogue, such as questions, statements, hesitations, etc. The knowledge of dialogue act realizations in a discourse or dialogue is part of the speech understanding and dialogue analysis process. It is of great importance for many applications: dialogue systems, speech recognition, automatic machine translation, etc. The main goal of this paper is to study the existing works about DA recognition and to discuss their respective advantages and drawbacks. A major concern in the DA recognition domain is that, although a few DA annotation schemes seem now to emerge as standards, most of the time, these DA tag-sets have to be adapted to the specificities of a given application, which prevents the deployment of standardized DA databases and evaluation procedures. The focus of this review is put on the various kinds of information that can be used to recognize DAs, such as prosody, lexical, etc., and on the types of models proposed so far to capture this information. Combining these information sources tends to appear nowadays as a prerequisite to recognize DAs.

Keywords: Bayesian approaches, dialogue act, lexical information, prosody, syntactic information

1 INTRODUCTION

Modeling and automatically identifying the structure of spontaneous dialogues is very important to better interpret and understand them. The precise modeling of spontaneous dialogues is still an open issue, but several specific characteristics of dialogues have already been clearly identified. Dialogue Acts (DAs) are one of these characteristics.

Austin defines in [1] the dialogue act as the meaning of an utterance at the level of illocutionary force. In other words, the dialogue act is the function of a sentence (or its part) in the dialogue. For example, the function of a question is to request some information, while an answer shall provide this information.

Dialogue acts can also be used in the context of Spoken Language Understanding. In such systems, dialogue acts are defined much more precisely, but are also application-dependent. Hence, Jeong et al. define in [2] a dialogue act as a domain-dependent intent, such as "Show Flight" or "Search Program" respectively in the flight reservation and electronic program guide domains.

Table 1 shows an example of the beginning of a dialogue between two friends, with Peter (A) calling Michal (B) on the phone. The corresponding DA labels are also shown. Each utterance is labeled with a unique DA.

Speaker	Dialogue Act	English
A	Conventional-opening	Hallo!?
B	Conventional-opening	Hi Peter!
B	Statement	It's me, Michael.
B	Question	How are you?
A	Conventional-opening	Hello Michael!
A	Statement	Very well.
A	Question	And you?
B	Statement	I'm well too.

Table 1. Example of the beginning of a dialogue between persons A and B in Czech, French and English with the corresponding DA labels.

1.1 Applications

There are many applications of automatic dialogue acts detection. We mention here only the most important ones: dialogue systems, machine translation, Automatic Speech Recognition (ASR), topic identification [3] and animation of talking head.

In dialogue systems, DAs can be used to recognize the intention of the user, for instance when the user is requesting some information and is waiting for it, or

when the system is trying to interpret the feedback from the user. An example of a dialogue management system that uses DA classification is the *VERBMOBIL* [4] system.

In machine translation, dialogue acts can be useful to choose the best solution when several translations are available. In particular, the grammatical form of an utterance may depend on its intention.

Automatic detection of dialogue acts can be used in ASR to increase the word recognition accuracy, as shown for example in [5]. In this work, a different language model is applied during recognition depending on the actual DA.

A talking head is a model of the human head that reproduces the speech of a speaker in real-time. It may also render facial expressions that are relevant to the current state of the discourse. Exploiting DA recognition in this context might make the animation more natural, for example by raising the eyebrows when a question is asked. Another easier option is to show this complementary information with symbols and colors near the head.

1.2 Objectives

Recognizing dialogue acts thus can be seen as the first level of dialogue understanding and is an important clue for applications, as it has been shown in the previous section. Several different dialogue act recognition approaches have been proposed in the literature. The main goal of this paper is to give a brief overview of these approaches. A short description is thus given for each of them, and is most often complemented by a discussion of their theoretical and practical advantages and drawbacks.

1.3 Paper Structure

This paper is organized as follows. The first section presents an introduction about the importance of dialogue act recognition with its main applications and objectives. Section 2 briefly describes the task of dialogue act recognition. Sections 3 and 4 describe the most common existing DA recognition approaches. The last section summarizes and discusses them altogether.

2 DIALOGUE ACT RECOGNITION

The first step to implement a dialogue act recognition system consists in defining the set of DAs labels that is relevant for the task. Then, informative features have to be computed from the speech signal and DA models are trained on these features. The segmentation of the dialogue into utterances may be carried out independently from DA recognition, or alternatively realized during the recognition step with joint DA recognition and segmentation models.

2.1 Dialogue Act Tag-set

The DA tag-set definition is an important but difficult step, because it results from a compromise between three conflicting requirements:

1. The DA labels should be generic enough to be useful for different tasks, or at least robust to the unpredictable variability and evolution of the target application;
2. The DA labels must be specific enough to encode detailed and exploitable characteristics of the target task;
3. The DA labels must be clear and easily separable, in order to maximize the agreement between human labelers.

Many different DA tag-sets can be found in the literature, the oldest being reviewed in [6]. Recently, a few of them seem to emerge as a common baseline, from which application-specific DA tags are derived. These are the Dialogue Act Markup in Several Layers (DAMSL) [7], the Switchboard SWBD-DAMSL [8], the Meeting Recorder [9], the VERBMOBIL [10] and the Map-Task [6] DAs tag-sets.

DAMSL was initially designed to be universal. Its annotation scheme is composed of four levels (or dimensions): communicative status, information level, forward looking functions and backward looking functions. Generally, these dimensions are considered as orthogonal and it shall be possible to build examples for any possible combination of them. The communicative status states whether the utterance is uninterpretable, abandoned or is a self-talk. This feature is not used for most of the utterances. The information level provides an abstract characterization of the content of the utterance. It is composed of four categories: task, task-management, communication-management and other-level. The forward looking functions are organized into a taxonomy, in a similar way as actions in traditional speech act theory. The backward looking functions show the relationship between the current utterance and the previous dialogue acts, such as accepting a proposal or answering the question. DAMSL is composed of 42 DA classes.

SWBD-DAMSL is the adaptation of DAMSL to the domain of telephone conversations. Most of the SWBD-DAMSL labels actually correspond to DAMSL labels. The Switchboard corpus utterances have first been labeled with 220 tags. 130 of those labels that occurred less than 10 times have been clustered, leading to 42 classes.

The Meeting Recorder DA (MRDA) tag-set is based on the SWBD-DAMSL taxonomy. The MRDA corpus contains about 72 hours of naturally occurring multiparty meetings manually-labeled with DAs and adjacency pairs. Meetings involve regions of high speaker overlap, affective variation, complicated interaction structures, abandoned or interrupted utterances, and other interesting turn-taking and discourse-level phenomena. The tags are not organized anymore on a dimensional level (such as DAMSL), but the correspondences are rather listed at the tag level. Each DA is described by one *general* tag, which may be for several DAs completed by one (or more) *specific* tag. A specific tag is used when the utterance cannot be

sufficiently characterised by a general tag only. For example, the utterance "Just write it down!" is characterised by the general tag *statement* and by the additional specific tag *command*. MRDA contains 11 general tags and 39 specific tags.

The DA hierarchy in VERBMOBIL is organized as a decision tree. This structure is chosen to facilitate the annotation process and to clarify relationships between different DAs. During the labeling process, the tree is parsed from the root to the leaves, and a decision about the next branch to parse is taken at each node (c.f. Figure 1).

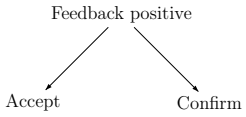

Fig. 1. Part of the VERBMOBIL DAs decision tree hierarchy.

42 DAs are defined in VERBMOBIL for German, English and Japanese, with 18 DAs at the illocutionary level.

The DA tags in the Map Task corpus [6] are structured into three levels, the highest modeling *transactions*, where each transaction accomplishes one major step in the speakers' plan. Transactions are then composed of *conversational games*, which model the regularity between questions/answers, statements/deny or acceptance, and so on. Games are finally made up of *conversational moves*, which classify different kinds of games according to their purposes. 19 moves are thus structured hierarchically into a decision-tree that is used to label each DA. For instance, the root of the trees splits into three moves: initiation, response and preparation. Initiation itself is then splitted into command, statement and question, and so on. Moves sequences are then delimited into conversational games, which start with an initiation and ends when that initiation's purpose is either fulfilled or abandoned. Each game is labeled with its purpose, whether it is a top level game or an embedded game, and is delimited in time. Transactions include task description, and are thus application-dependent.

2.2 Dialogue Act Recognition Information

The most important types of information commonly used to recognize dialogue acts are described below.

The first one is **lexical information**. Every utterance is composed of a sequence of words. Generally, the DA of an utterance can be partly deduced from the lists of words that form this utterance. For example, Wh-questions often contains an interrogative word, which rarely occurs in other DA classes. Lexical information is typically captured by words unigrams.

The second one is **syntactic information**. It is related to the *order* of the words in the utterance. For instance, in French and Czech, the relative order of the *subject* and *verb* occurrences might be used to discriminate between declarations and questions. Words n-grams are often used in dialogue act recognition to model some local syntactic information. Král et al. propose in [11] to further model words position in the utterance in order to also take into account global syntactic information. Another type of syntactic information recently used for DA recognition are "cue phrases", which actually corresponds to a subset of specific n-grams, where n may vary from 1 to 4, which are selected based on their capacity to predict a specific dialogue act and on their occurrence frequency [12]. These cue phrases actually correspond to common and typical sequences of words. As they do not model the whole lexical space, one might interpret them in a context of DA detection instead of DA recognition.

Another information is **semantic information**. The DA also depends on the meanings of the utterance and the words that compose it. However, many different definitions of "semantic information" exist, ranging from broad topic categories such as "weather", "sports", down to precise frame-based interpretations, e.g. "show flights from London to Paris on march 12th". The latter is typically used in spoken language understanding applications, where a dialogue act is dependent on a specific pre-defined action [2]. Another kind of semantic information that is used in DA recognition is specific entities, such as named or task entities. For instance, date, place or proper nouns, when they are uttered, may constitute important cues to find out what is the utterance dialogue act [13]. Also, Bangalore et al. use in [14] speaker and task entities as features. They obtain a DA error rate of 38.8% with 67 dialogue acts adapted from DAMSL on a product ordering task.

Yet another useful information to recognize DAs is **prosody**, and more particularly the melody of the utterance. Usually, questions have an increasing melody at the end of utterance, while statements are often characterised by a slightly decreasing melody.

The last information mentioned here is the **context** of each DA. Hence, any DA depends on the previous (and next) DAs, the most important context being the previous one. For example, a "Yes" or "No" answer is most likely to follow a *Yes/no question*. The sequence of DAs is also called the *dialogue history*.

We focus next on the three following information sources, which are the most commonly used in application-independent DA recognition systems [15, 16]:

- Lexical (and syntactic) information
- Prosodic information
- Dialogue history

2.2.1 Lexical Information

Lexical and syntactic features can be derived from the word sequence in the dialogue. The first broad group of DA recognition approaches that uses this type of features is based on the assumption that different dialogue acts are generally composed of sequences of different words.

The correspondence between DAs and words sequences is usually represented either by n-grams, Naive Bayes, Hidden Markov Models, Bayesian Networks, etc. (see section 3), or Non-Bayesian approaches, such as Neural Networks, Semantic Classification and Regression Trees, etc. (see section 4).

2.2.2 Prosodic Information

Most researchers agree on the fact that the lexical/syntactic information is not generally sufficient to explain DAs. Prosodic cues [17] are also related to DA instances.

For example, questions are usually characterized by an increasing melody at the end of the utterance [18], and accepts have usually much more energy than backchannels and acknowledgments [9].

Prosody is successfully used in [19] for French and Vietnamese question detection. Authors exploit the fact that French questions are usually characterized by their intonation curves. The set of prosodic features is derived from the curve of the fundamental frequency (F0). Some features are F0 statistics (Min, Max, Mean, etc.), while other features describe whether F0 is raising or falling. According to the authors, Vietnamese questions and affirmative sentence differ in the F0 contour at the final segment of the sentence, both in register and intensity. They respectively obtain on the French DELOC (telephone meetings) and NESPOLE [20] corpora 74% and 73% of accuraccy. Their question detection accuracy on the Vietnamese VietP corpus is 77%.

Prosodic features are usually modeled with the same Bayesian or Non-Bayesian methods as used for lexical information.

2.2.3 Dialogue History

The third general type of information used in classical DA recognition systems is the dialogue history. It is defined by the sequence of previous DAs that have been recognized. It may be used to predict the next DA. Different formalisms are employed to model this information: statistical models such as n-grams, Hidden Markov Models (HMMs), Bayesian Networks, etc.

2.3 Segmentation

To recognize DAs, the dialogue must first be segmented into sentence-level units, or utterances [21], where each utterance represents a single DA. Segmentation of the dialogue into such utterances may be carried out separately or realized during the recognition step.

The hidden-event language model has been proposed in [22] to automatically detect utterance boundaries. Its basic principle consists in modeling the joint probability of words and sentence boundaries with an n-gram. The training of the model is realized as in the classical n-gram case with a new token that represents the DA boundary. Shriberg et al. show in [23] that prosodic features give better results than lexical features to segment utterances.

Kolář et al. show in [24] an extension of this approach. They adapt the hidden-event language models to the speaker to improve dialogue act segmentation accuracy. Speaker adaptation is realized by linear combination of the speaker independent and speaker dependent language models. They use ICSI meeting corpus [25].

Ang et al. use in [26] a decision tree that estimates the probability of occurrence of a DA boundary after each word based on the length of the pause between contiguous words of the same speaker, and a bagging classifier that models prosodic attributes. This approach is further combined via an HMM with an hidden-event language model.

The main focus of this review being dialogue act recognition, we will not detail more in the following the works about utterance segmentation. Please refer for example to [27] for an overview of this domain.

3 BAYESIAN APPROACHES

The main types of automatic DA recognition approaches proposed in the literature can be broadly classified into Bayesian and Non-Bayesian approaches. Bayesian approaches are presented in this section and Non-Bayesian approaches are described in Section 4.

3.1 Lexical (and Syntactic) N-Gram DA Models

The Bayesian formalism has been the preferred approach in the DA recognition domain for a long time now. For instance, [28] finds the best sequence of dialogue acts $\hat{C}$ by maximizing the *a posteriori* probability $P(C|O)$ over all possible sequences of dialogue acts C as follows:

$$\begin{aligned} \hat{C} &= \arg\max_C P(C|O) \\ &= \arg\max_C \frac{P(C).P(O|C)}{P(O)} \\ &= \arg\max_C P(C).P(O|C) \end{aligned} \tag{1}$$

The most common methods model $P(O|C) = P(W|C)$, where W is the word sequence in the pronounced utterance with statistic models such as n-grams. These methods are based on the observation that different DA classes are composed of distinctive word strings. For example, 92.4% of the "uh-huh" occur in Backchannels

and 88.4% of the trigrams "<start> do you" occur in yes-no questions [15]. The words order and positions in the utterance may also be considered. A theory of word frequencies, which is the basis for DA modeling from word features, is described in [3].

3.1.1 DA Recognition from Exact Words Transcriptions

The following approach is based on the hypothesis that the words in the utterances are known. Then, Equation 1 becomes:

$$\arg\max_C P(C|W) = \arg\max_C P(C).P(W|C) \tag{2}$$

The "Naive Bayes assumption", which assumes independence between successive words, can be applied and leads to:

$$\arg\max_C P(C).P(W|C) = \arg\max_C P(C).\prod_{i=1}^{T} P(w_i|C) \tag{3}$$

This equation represents the unigram model, also sometimes called the Naive Bayes classifier. In this case, only lexical information is used. Higher order models, such as 2-grams, 3-grams, etc., also take into account some local syntactic information about the dependencies between adjacent words. Because of limited corpus sizes, the use of 4-grams and more complex models is rare.

Reithinger et al. use in [29] unigram and bigram language models for DA recognition on the VERBMOBIL corpus. Their DA recognition rate is about 66% for German and 74% for English with 18 dialogue acts. In [30], a naive Bayes n-gram classifier is applied to the English and German language. The authors obtain a DA recognition rate of 51% for English and 46% for German on the NESPOLE corpus. Grau et al. use in [31] the naive Bayes and uniform naive Bayes classifiers with 3-grams. Different smoothing methods (Laplace and Witten Bell) are evaluated. The obtained recognition rate is 66% on the SWBD-DAMSL corpus with 42 DAs. Ivanovic also uses in [32] the naive Bayes n-grams classifier and obtains about 80% of recognition rate in the instant messaging chat sessions domain with 12 DAs classes derived from the 42 DAs of DAMSL.

One can further assume that all DA classes are equi-probable, and thus leave the $P(C)$ term out:

$$\hat{C} = \arg\max_C P(W|C) \tag{4}$$

This approach is referred to as the *uniform* naive Bayes classifier in [31].

3.1.2 DA Recognition from Automatic Word Transcription

In many real applications, the exact words transcription is not known. It can be approximately computed from the outputs of an automatic speech recognizer. Let A

be a random variable that represents the acoustic information of the speech stream (e.g. spectral features).

The word sequence W is now an hidden variable, and the observation likelihood $P(A|C)$ can be computed as:

$$P(A|C) = \sum_W P(A|W,C).P(W|C) \quad (5)$$

$$= \sum_W P(A|W).P(W|C) \quad (6)$$

where C is the DA class and $P(A|W)$ is the observation likelihood computed by the speech recognizer for a given hypothesized word sequence W. Most of the works on Bayesian dialogue act recognition from speech, such as in [15], use this approach and approximate the summation over the k-best words sequence only.

3.2 Dialogue Sequence N-Gram Models

The dialogue history also contains very important information to predict the current DA based on the previous ones. The dialogue history is usually modeled by a statistical discourse grammar, which represents the prior probability $P(C)$ of a DA sequence C.

Let C_τ be a random variable that represents the current dialogue act class at time τ. The dialogue history H is defined as the previous sequence of DAs: $H = (C_1, .., C_{\tau-1})$. It is usually reduced to the most recent n DAs: $H = (C_{\tau-n+1}, .., C_{\tau-1})$. The most common values for n are 2 and 3, leading to 2-gram and 3-gram models. In order to train such models, the conditional probabilities $P(C_\tau|C_{\tau-n+1}, .., C_{\tau-1})$ are computed on a labeled training corpus. *Smoothing* techniques, such as standard back-off methods [33], may also be used to train high-order n-grams. N-grams are successfully used to model dialogue history in [15, 34].

Polygrams are mixtures of n-grams of varying order: n can be chosen arbitrarily large and the probabilities of higher order n-grams are interpolated by lower order ones. They usually give better recognition accuracy than standard n-grams and are shown in [35].

3.3 Hidden Markov Models

Hidden Markov Models can also be used as in [15] to model sequences of dialogue acts. Let O be a random variable that represents the observations and C the sequence of DAs classes. n th-order HMM can be considered, which means that each dialogue act depends on the n previous DAs (in a similar way as for n-grams). Then, each HMM state models one DA and the observations correspond to utterance level features. The transition probabilities are trained on a DA-labeled training corpus.

DA recognition is carried out using some dynamic programming algorithm such as the Viterbi algorithm.

HMMs with word-based and prosodic features are successfully used to model dialogue history in [36]. [5] uses intonation events and tilt features such as: F0 (fall/rise, etc.), energy, duration, etc. She achieves 64% of accuracy on the DCIEM map task corpus [37] with 12 DA classes. Ries combines in [38] HMMs with neural networks (c.f. Section 4.1). He obtains about 76% of accuracy on the CallHome Spanish corpus. In [39] language models and modified HMMs are applied on the Switchboard corpus [40] with the SWBD-DAMSL tag-set.

3.4 Bayesian Networks

A Bayesian network is represented by a directed acyclic graph. Nodes and arcs respectively represent random variables and relations (dependencies) between nodes. The topology of the graph models conditional independencies between the random variables. In the following, we do not differentiate dynamic Bayesian networks (with stochastic variables) from static Bayesian networks, as most of our variables are stochastic, and when static Bayesian networks are drawn, they represent an excerpt of a dynamic Bayesian network at a given time slice. The stochastic variables are conditionally dependent of theirs descendants and independent of theirs ascendants.

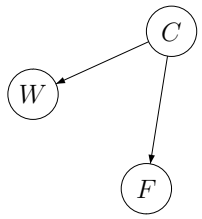

Fig. 2. Example of Bayesian network for dialogue act recognition.

An example of Bayesian network for dialogue act recognition is shown in Figure 2. Node C represents the current dialogue act. Utterance features are represented by nodes W (sequence of words in the utterance) and F (prosodic features). The dialogue context is not considered there. The conditional independence assertions of this network allows the following factorization:

$$P(C, W, F) = P(W|C).P(F|C).P(C) \tag{7}$$

In order to build such a network, the network structure (conditional dependencies) and the conditional probability distributions must be defined. The conditional probabilities are trained statistically on a training corpus. The topology of network can be created manually or automatically.

Bayesian networks are successfully used in [41] for dialogue act recognition. In the first experiment reported, three features are used: sentence type (declarative, yes/no question, etc.), subject type (1st/2nd/3rd person) and punctuation (question mark, exclamation mark, comma, etc). The Bayesian network is defined manually. They achieve 44% of accuracy on the SCHISMA corpus [42]. In the second experiment, a small corpus is derived from the dialogue system used to interact with the navigation agent. Utterances are described by surface level features, mainly keywords-based features. These features are computed automatically for each utterance. Bayesian networks are further generated automatically iteratively, starting from a small hand-labeled DA corpus. This network is used to parse another large corpus, and a new network is generated from this corpus. This approach gives 77% of accuracy for classification of forward-looking functions (7 classes) and 88% of accuracy for backward-looking functions (3 classes).

Another application of Bayesian network in dialogue act recognition is shown in [43]. Two types of features are used: utterance features (words in the utterance: w_i) and context features (previous dialogue act: $C_{\tau-1}$). The authors compare two different Bayesian networks to recognize DAs (see Figure 3).

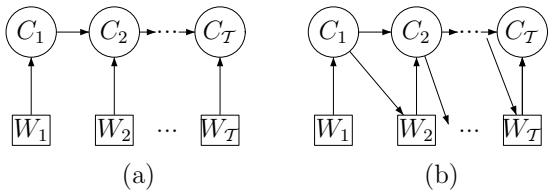

Fig. 3. Two Bayesian networks for dialogue act recognition: C_i represents a single DA, while W_i is a sequence of words.

These networks are built manually. In the left model of Figure 3, each dialogue act is recognized from the words of the current utterance and from the previous DA. In the right model of Figure 3, the authors further consider an additional dependency between each word of the utterance and its previous dialogue act (diagonal arcs). They achieve about 64% precision on a subset of the MRDA corpus and with the reduced DA set size.

Another Bayesian model, the triangular-chain conditional random field, which jointly models dialogue acts and named entities, has been proposed in [2]. This model is shown in Figure 4.

This joint model is shown to outperform sequential and cascade models, in which dialogue acts are assumed independent of named entities. In the independent approach, DAs are often modeled by a multivariate logistic regression (or maximum

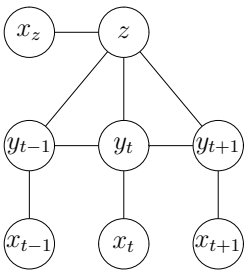

Fig. 4. Triangular-chain Conditional Random Field, from [2]. It is used to jointly models dialogue acts (represented by variables z) and named entities (represented by variables y). Variable x encodes the words sequence.

entropy classifier)

$$p(z|x) = \frac{1}{Z_z(x)} \exp\left(\sum_k \nu_k h_k(z, x)\right)$$

that maximizes the entropy $h_k(z, x)$, where z is the DA and x the words sequence. Alternatively, the joint model combines both maximum entropy and conditional random fields approaches.

Dynamic Bayesian Network (DBN) have also successfully been used for DA recognition in [44], where a switching DBN combines several partial models and coordinates the DA recognition task. The relation between the sequences of transcribed words and their DA labels is modeled by an interpolated Factored Language Model (FLM), while the dialogue history is represented by a trigram language model. Prosodic features (pitch, energy, etc.) are also used for segmentation. The proposed approach is based on a switching DBN model that alternates between two submodels: an *intra-DA model* that represents a single DA class associated to a words sequence, and an *inter-DA model* that is activated at DA boundaries. A dedicated random variable of these models is used to detect these DA boundaries. The authors obtain on the AMI Meeting Corpus [45] about 60% of DA tagging rate with 15 DA classes.

4 NON-BAYESIAN APPROACHES

Non-Bayesian approaches are also successfully used in the DA recognition domain, but they are not so popular as Bayesian approaches. Examples of such approaches are Neural Networks (NNs), such as Multi-Layer Perceptron (MLP) or Kohonen Networks, Decision Trees, Memory-Based Learning and Transformation-Based Learning.

4.1 Neural Networks

A neural network (NN) [46] is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing based on a connectionist approach to computation. It can be used to model complex relationships between inputs and outputs or to find patterns in data.

4.1.1 Multi-Layer Perceptron

One of the most frequently used neural network technique in the DA recognition domain is the multi-layer perceptron (MLP, see Figure 5), which consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and one output layer. The input signal propagates through the network layer-by-layer. An MLP can represent a non linear function.

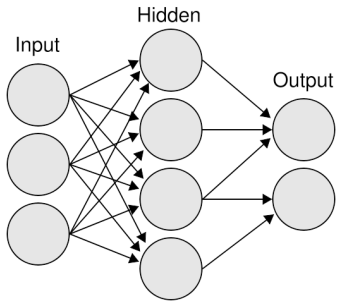

Fig. 5. Example of multi-layer perceptron.

Wright describes in [5] an approach with a one-hidden-layer MLP. 54 suprasegmental and duration prosodic features are used as inputs. She achieves 62% of accuracy on the DCIEM map task corpus [37] with 12 DA classes. Ries successfully uses in [38] an MLP both stand-alone, and in combination with HMMs. He obtains a similar accuracy (about 76%) on the CallHome Spanish corpus with both setups. Sanchis et al. also use in [47] an MLP to recognize DAs. The features considered are the words of the lexicon restricted to the semantic task (138 inputs=size of the lexicon). The experiments are performed on the Spanish dialogue corpus in the train transport domain (16 DA classes). They achieve about 93% of accuracy on the text data and about 72% of accuracy on the recognized speech. Note that this approach may be difficult to apply on a large lexicon. Levin et al. use in [30] a set of binary features to train an MLP. These features are computed automatically by combining grammar-based phrasal parsing and machine learning techniques. They obtain a DA recognition accuracy of about 71% for English and about 69% for German on the NESPOLE corpus.

4.1.2 Kohonen Networks

Another type of neural network used in the dialogue act classification domain is Kohonen Networks. A Kohonen network [48], also known as Self-Organizing Map (SOM), defines an ordered mapping, a kind of projection from a set of given data items onto a regular, usually two-dimensional grid. A model is associated with each grid node (see Figure 6).

The topology of the SOMs is a single layer feedforward network where the discrete outputs are arranged into a low dimensional (usually 2D or 3D) grid. Each input is connected to all output neurons. A weight vector with the same dimensionality as the input vectors is attached to every neuron. The number of input dimensions is usually much larger than the output grid dimension. SOMs are mainly used for dimensionality reduction.

The models of the Kohonen network are estimated by the SOM algorithm [49]. A data item is mapped onto the node which model is the most similar to the data item, i.e. has the smallest distance to the data item, based on some metric.

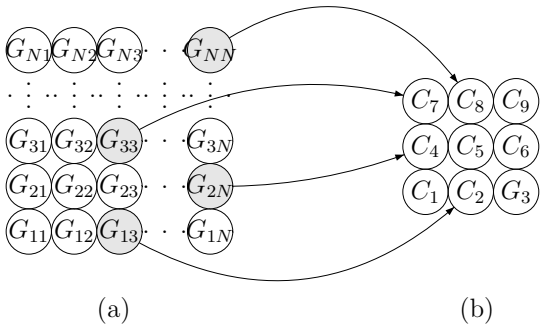

Fig. 6. Two Kohonen networks (from [50]) with a rectangular structure to model dialogue acts: The inputs to the large network (on the left) are a set of binary utterance features. Neurons representative of DA classes are grayed. The small network on the right represents the outputs of system (DA classes). The connexions between the neighboring nodes are not shown.

Kohonen networks for dialogue act recognition are used in [50]. The authors use seven *superficial* utterance features: speaker, sentence mode, presence or absence of a wh-word, presence or absence of a question mark, etc. Each utterance is represented by a pattern of these features, which is encoded into a binary format for the SOM representation. Initially, the exact number of DA classes is not known *a priori*, and only the large network on the left is created and trained. The clustering process is interrupted after a given number of clusters have been found.

To interpret the clusters, another small Kohonen network is built (right model in Figure 6). This network contains as many neurons as DA classes. These neurons

are initialized by the values of the weight-vectors of the representative neurons from the large network.

The quality of classification is evaluated by the Specificity Index (SI) [51] and by the Mean number of Conditions (MoC). They achieve about 0.1 for SI and about 2.6 for MoC on the SCHISMA corpus, with 15 DA classes and a network with 10*10 neurons. Another experiment has been performed with 16 DA classes and a larger network with 12*12 neurons with comparable results. Generally, unsupervised methods such as Kohonen networks are rarely used for DA recognition.

4.2 Decision Trees

Decision trees (or Classification and Regression Trees, CARTs) [52] are generation tools that are successfully used in operations research and decision analysis. They are usually represented by an oriented acyclic graph (see Figure 7). The root of the tree represents the starting point of the decision, each node contains a set of conditions to evaluate, and arcs show the possible outcomes of these decisions.

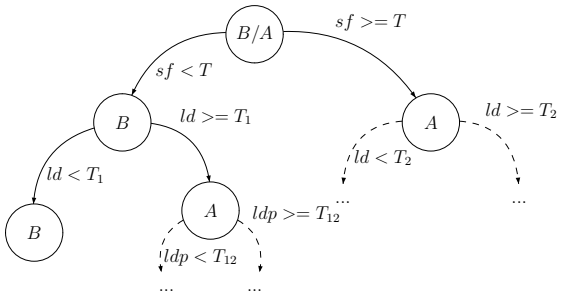

Fig. 7. Example of a part of the decision tree in the DA recognition domain: recognition of Backchannels (B) and Accepts (A) by prosody, from [16].

In the case of DA recognition, the decisions usually concern utterance features. Each decision compares the value of some feature with a threshold. For example, in Figure 7, three different prosodic features (sf, ld and ldp) are shown with their corresponding thresholds (T, T_1, T_2 and T_{12}). sf is the pause type feature and ld and ldp are the duration type features. Training of the decision tree is performed automatically on the training corpus. The output of the CART is the probability of the DA given the utterance features (lexical and prosodic), i.e., the *posterior* probability $P(C|W, F)$. The main advantage of CARTs is that they can combine different discrete and continuous features.

Wright uses in [5] 54 suprasegmental and duration prosodic features to train the trees on the CART algorithm [52]. She achieves 63% of accuracy on the DCIEM map task corpus with 12 DA classes. Shriberg et al. also use in [16] CARTs for DA recognition with prosodic features. They use CARTs to recognize a few DAs only,

which are very difficult to recognize with lexical (and syntactic) features. These DAs are recognized from prosody only. CARTs are used for example to distinguish statements from questions because questions usually differ from statements by an increasing final F0 curve. Therefore, this CART classifier is trained on statements and questions data only. Levin et al. compare in [30] CARTs with other classifiers, mainly Naive Bayes and MLP classifiers. They use binary grammatical features for this comparison. They show that CARTs outperform the Naive Bayes classifier and that they give comparable results with an MLP. The resulting DA recognition accuracy is about 68% for English and about 66% for German on the NESPOLE corpus.

4.3 Memory-Based Learning

Memory-Based Learning (MBL) [53] is an application of the memory-based reasoning theory in the field of machine learning. This theory is based on the assumption that it is possible to handle a new sample by matching it with stored representations of previous samples. Hence, in MBL, all known samples are stored in memory for future reference, and any unknown sample is classified by comparing it with all the stored samples. The main advantage of MBL compared to other machine learning techniques is that it successfully manages exceptions and sub-regularities in data. The main drawback of the method is its high memory and computational requirements.

Several methods can be used to compare the stored and recognized samples. The most popular one is the k-Nearest Neighbor (k-NN) [54]. It consists in defining a distance measure between samples, and of retrieving the k stored samples that have the smallest distance to the target one. These k samples are assumed to be similar to the recognized one, and the recognized sample is classified into the dominant class amongst these "neighbors".

Rotaru uses in [55] MBLs in an automatic dialogue acts tagging task on the Switchboard corpus [40] of spontaneous human-human telephone speech. The utterance features are based on word bigrams computed on the whole training corpus. These bigrams are hashed to a given number of features, which optimal value is found experimentally. The hash function uses the letters present in the bigrams and the number of features. The author experiments a various number of neighbors. The best performance is about 72% of accuracy with three neighbors. Levin et al. exploit in [30] MBLs on the NESPOLE corpus. They use the same features as described in the MLP case (see Section 4.1.1) on the IB1 algorithm [56] with one neighbor. They achieve about 70% of accuracy for English and about 67% for German. MBLs are also used in [57] with the IB1 algorithm. The authors obtain an accuracy of about 74% with prosodic, lexical and context features on a corpus of Dutch telephone dialogues between users and the Dutch train timetable information system.

4.4 Transformation-Based Learning

The main idea of Transformation-Based Learning (TBL) [58] is to start from some simple solution to the problem, and to apply transformations to obtain the final result. Transformations are composed in a supervised way. Given a labeled training corpus and a set of possible transformation templates on this corpus, all possible transformations are generated from the templates, after what the transformations are selected iteratively. The templates can be for example: if tag X is after tag Y and/or N previous utterances contain word w, then change actual tag to Z. At each step the "best" transformation (bringing the largest improvement to precision) is selected and applied to the current solution. The algorithm stops when the selected transformation does not modify the data enough, or when there are no more transformations left.

The total number of all possible transformations can be very high. It is thus often computationally expensive to test all transformations, especially since most of them do not improve precision. A Monte-Carlo (MC) approach [59] can be used to tackle this issue: only a fixed number of transformations are selected randomly and used in the next steps. Although this may exclude the best transformation from the retained set, there are usually enough transformations left so that one of them still brings a large improvement to precision.

TBL can be applied to most classification tasks, and has been proposed for automatic DA recognition and some related works. [60] use TBL with a Monte Carlo strategy on the VERBMOBIL corpus. They use the following utterance features for DA recognition: cue phrases, word n-grams, speaker identity, punctuation marks, the preceding dialogue act, etc. The resulting DA accuracy is about 71% with 18 dialogue acts. Bosch et al. use in [61] TBLs on the corpus of Dutch telephone dialogues between users and the Dutch train timetable information system, with a very limited DA tag-set. Question-answer pairs are represented by the following feature vectors: six features represent the history of questions asked by the system, while the following features represent the recognized user utterance, which is encoded as a sequence of bits, with 1 indicating that the *i-th* word of the lexicon occurs at least one time in the word graph. The last feature is used for each user utterance to indicate whether this sentence gave rise to a communication problem or not, as requested by the application, which final objective is to detect communication problems (incorrect system understanding) between the user and the dialogue system. They achieve to detect about 91% of all communication problems with the rule-induction algorithm RIPPER [62]. The authors show that TBL outperforms MBL on this task. Lendvai et al. also use in [57] TBLs with the RIPPER algorithm. They obtain an accuracy of about 60% with prosodic, lexical and context features on the same Dutch corpus as in the previous experiments.

4.5 Meta-Models

Model probabilities, such as the ones computed by the lexical n-gram previously described, can also be used as features of a "meta-model", which role is to combine different sources of information in order to disambiguate the utterance. Hidden Markov Models are typically used for this purpose, as already described in section 3.3. Another solution exploits boosting and committee-based sampling techniques, which can be used to compute tagging confidence measures, such as in [60], or to recognize sub-tasks labels [63], where a sub-task is defined as a sequence of DAs. Zimmermann compares in [64] n-gram, cue-phrases, maximum entropy and boosting classifiers for dialogue act recognition on a meeting corpus. On the ICSI MRDA meeting corpus, they obtain 23.3% of DA recognition accuracy with 5 DA classes, by combining four individual DA classifiers: n-grams, cue phrases, maximum entropy and boosting. Combination is realized with an MLP.

5 DISCUSSION AND CONCLUSIONS

Automatic recognition of dialogue acts is an important yet still underestimated component of Human-Machine Interaction dialogue architectures. As shown in this review, research in this area have made great progresses during the last years. Hence a few DA tag-sets have emerged as pseudo-standards and are more and more often used in the community. Nevertheless, these tag-sets are nearly always manually adapted to fit the specificities of each particular application, which points out a major issue in this area that concerns the variability of dialogue acts definitions and the consequent excessive costs and difficulty to port some previous work to a new task.

Another interesting characteristic of the dialogue act recognition domain is the fact that several different sources of information have to be combined to achieve reasonably good performances. In particular, most of the works discussed in this review show the importance to combine both lexical and prosodic information, as well as higher-level knowledge such as the overall structure of the dialogue used in the task, or semantic information such as named or task-related entities. This confirms our intuition that dialogue act recognition is a rich research area that might benefit from a better understanding of the dialogue processing, in particular with regard to the context of the dialog. Hence, many contextual relevant information are still not considered, for instance the social roles and relationships between the users, the emotions of the speakers, the surrounding environment as well as the past and recent history of interaction. All these information considerably influence the course of a dialogue, but are also extremely difficult to model and thus to include in our models. However, we have seen that the domain has progressively seen its influence area grows and intersects more and more with other research areas: from text to speech, from lexicon to prosody and semantic. We are convinced that this progression should continue, and that the overlap with adjacent domains should keep on enlarging, which is easier to achieve now thanks to the recent progresses

realized in, for example, the fields of user modeling and collaborative filtering, or emotions recognition, just to cite a few.

ACKNOWLEDGMENT

This work has been partly supported by the Ministry of Education, Youth and Sports of Czech Republic grant (NPV II-2C06009).

*

REFERENCES

[1] J. L. Austin. How to do Things with Words. *Clarendon Press, Oxford*, 1962.

[2] M. Jeong and G. G. Lee. Jointly predicting dialog act and named entity for spoken language understanding. 2006.

[3] P. N. Garner, S. R. Browning, R. K. Moore, and R. J. Russel. A Theory of Word Frequencies and its Application to Dialogue Move Recognition. In *ICSLP'96*, volume 3, pages 1880–1883, Philadelphia, USA, 1996.

[4] Jan Alexandersson, Norbert Reithinger, and Elisabeth Maier. Insights into the Dialogue Processing of VERBMOBIL. Technical Report 191, Saarbrücken, Germany, 1997.

[5] H. Wright. Automatic Utterance Type Detection Using Suprasegmental Features. In *ICSLP'98*, volume 4, page 1403, Sydney, Australia, 1998.

[6] J. Carletta, A. Isard, S. Isard, J. Kowtko, A. Newlands, G. Doherty-Sneddon, and A. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23:13–31, 1997.

[7] J. Allen and M. Core. Draft of Damsl: Dialog Act Markup in Several Layers. In *http://www.cs.rochester.edu/research/cisd/resources/damsl/ RevisedManual/RevisedManual.html*, 1997.

[8] D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation (Coders Manual, Draft 13). Technical Report 97-01, University of Colorado, Institute of Cognitive Science, 1997.

[9] R. Dhillon, Bhagat S., H. Carvey, and Shriberg E. Meeting Recorder Project: Dialog Act Labeling Guide. Technical Report TR-04-002, International Computer Science Institute, February 9 2004.

[10] S. Jekat *et al.* Dialogue Acts in VERBMOBIL. In *Verbmobil Report 65*, 1995.

[11] P. Král, C. Cerisara, and J. Klečková. Lexical Structure for Dialogue Act Recognition. *Journal of Multimedia (JMM)*, 2(3):1–8, June 2007.

[12] N. Webb, M. Hepple, and Y. Wilks. Dialog act classification based on intra-utterance features. Technical Report CS-05-01, Dept of Comp. Science, University of Sheffield, 2005.

[13] S. Rosset, D. Tribout, and L. Lamel. Multi-level information and automatic dialog act detection in human-human spoken dialogs. *Speech Communication*, 2008.

[14] S. Bangalore, G. Di Fabbrizio, and A. Stent. Towards Learning to Converse: Structuring Task-oriented Human-human Dialogs. In *ICASSP'06*, Toulouse, France, May 2006.

[15] A. Stolcke *et al.* Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech. In *Computational Linguistics*, volume 26, pages 339–373, 2000.

[16] E. Shriberg *et al.* Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? In *Language and Speech*, volume 41, pages 439–487, 1998.

[17] P. Langlais. *Traitement de la prosodie en reconnaissance automatique de la parole.* PhD thesis, Univerité d'Avignon et des pays de Vaucluse, 1995.

[18] P. Martin. Prosodic and Rhythmic Structures in French. In *Linguistics*, volume II, pages 925–949, 1987.

[19] V. M. Quang, L. Besacier, and Castelli E. Automatic Question Detection Prosodic-lexical Features and Crosslingual Experiments. In *Interspeech'2007*, pages 2257–2260, Antwerp, Belgium, August, 27-31 2007.

[20] N. Mana *et al.* The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains. In *Eurospeech'2003*, Geneva, Switzerland, September, 1-4 2003.

[21] M. Meteer, A. Taylor, R. MacIntyre, and R. Iyer. Dysfluency Annotation Stylebook for the Switchboard Corpus. Technical report, Linguistic Data Consortium, Ferbruary 1995. ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps, Revised June 1995 by A. Taylor.

[22] A. Stolcke and E. Shriberg. Automatic linguistic segmentation of conversational speech. In *Proc. ICSLP '96*, volume 2, pages 1005–1008, Philadelphia, PA, 1996.

[23] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. In *Speech communication*, volume 32, pages 127–154, September 2000.

[24] J. Kolar, Y. Liu, and E. Shriberg. Speaker adaptation of language models for automatic dialog act segmentation of meetings. In *Interspeech'2007*, pages 1621–1624, Antwerp, Belgium, August, 27-31 2007.

[25] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The ICSI Meeting Corpus. In *ICASSP'2003*, pages 364–367, Hong Kong, April 2003.

[26] J. Ang, Y. Liu, and E. Shriberg. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. In *Proc. ICASSP*, March 2005.

[27] Yang Liu. *Structural event detection for rich transcription of speech.* PhD thesis, Purdue University, December 2004.

[28] J. O. Berger. Statistical Decision Theory and Bayesian Analysis. *Springer-Verlag, New York*, 1985.

[29] N. Reithinger and M. Klesen. Dialogue Act Classification Using Language Models. In *EuroSpeech'97*, pages 2235–2238, Rhodes, Greece, September 1997.

[30] L. Levin, C. Langley, A. Lavie, D. Gates, D. Wallace, and K. Peterson. Domain Specific Speech Acts for Spoken Language Translation. In *4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, 2003.

[31] S. Grau, E. Sanchis, M. J. Castro, and D. Vilar. Dialogue Act Classification using a Bayesian Approach. In *9th International Conference Speech and Computer (SPECOM'2004)*, pages 495–499, Saint-Petersburg, Russia, September 2004.

[32] E. Ivanovic. Dialogue Act Tagging for Instant Messaging Chat Sessions. In *ACL Student Research Workshop*, pages 79–84, Ann Arbor, Michigan, USA, June 2005. Association for Computational Linguistics.

[33] J. Bilmes and K. Kirchhoff. Factored Language Models and Generalized Parallel Backoff. In *Human Language Technology Conference*, Edmonton, Canada, 2003.

[34] N. Reithinger and E. Maier. Utilizing Statistical Dialogue Act Processing in VERBMOBIL. In *33rd annual meeting on Association for Computational Linguistics*, pages 116–121, Morristown, NJ, USA, 1995. Association for Computational Linguistics.

[35] M. Mast *et al.* Automatic Classification of Dialog Acts with Semantic Classification Trees and Polygrams. In *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 217–229, 1996.

[36] A. Stolcke *et al.* Dialog Act Modeling for Conversational Speech. In *AAAI Spring Symp. on Appl. Machine Learning to Discourse Processing*, pages 98–105, 1998.

[37] E. G. Bard, C. Sotillo, A. H. Anderson, and M. M. Taylor. The DCIEM Map Task Corpus: Spontaneous Dialogue Under Sleep Deprivation and Drug Treatment. In *ICSLP'96*, volume 3, pages 1958–1961, Philadelphia, USA, 1996.

[38] K. Ries. HMM and Neural Network Based Speech Act Detection. In *ICASSP'99*, volume 3, pages 497–500, 1999.

[39] P. T. Douglas, , and F. N. Jay. Speech Act Profiling: A Probabilistic Method for Analyzing Persistent Conversations and their Participants. In *37th Annual Hawaii International Conference on System Sciences (HICSS'04)*. IEEE, 2004.

[40] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP'1992*, volume 1, pages 517–520, San Francisco, CA, USA, 23-26 March 1992.

[41] S. Keizer, Akker. R., and A. Nijholt. Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues. In *3rd ACL/SIGdial Workshop on Discourse and Dialogue*, pages 88–94, Philadelphia, USA, July 2002.

[42] Simon Keizer. Dialogue Act Classification: Experiments with the SCHISMA Corpus. Technical report, University of Twente, October 2002.

[43] G. Ji and J. Bilmes. Dialog Act Tagging Using Graphical Models. In *ICASSP'05*, volume 1, pages 33–36, Philadelphia, USA, March 2005.

[44] A. Dielmann and S. Renals. DBN Based Joint Dialogue Act Recognition of Multiparty Meetings. In *ICASSP'07*, pages 133–136, Honolulu, Hawaii, USA, April 2007.

[45] J. Carletta *et al.* The AMI Meeting Corpus: A Preannouncement. In *Multimodal Interaction and Related Machine Learning Algorithm Workshop (MLMI-05)*, Edinburgh, UK, July, 11-13 2005.

[46] S. Haykin. *Neural Networks: a Comprehesive Foundation*. Prentice Hall, 2nd edition, 1999.

[47] E. Sanchis and M. J. Castro. Dialogue Act Connectionist Detection in a Spoken Dialogue System. In *Second International Conference on Hybrid Intelligent Systems (HIS2002)*, pages 644–651, Santiago de Chile, Chile, 1-4 December 2002. IOS Press.

[48] T. Kohonen. Self-Organizing Maps. *Springer Series in Information Sciences*, 30, 1995.

[49] M. Cottrell and J.C. Fort. Theoretical Aspects of the SOM Algorithm. *Neurocomputing*, 21:119–138, 1998.

[50] T. Andernach, M. Poel, and E. Salomons. Finding Classes of Dialogue Utterances with Kohonen Networks. In *ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pages 85–94, Prague, Czech Republic, April 1997.

[51] T. Andernach. A Machine Learning Approach to the Classification of Dialogue Utterances. In *NeMLaP-2*, Ankara, Turkey, July 1996.

[52] L. Breiman, J. Friedman, and R. Olshen. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA, 1984.

[53] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. TiMBL: Tilburg Memory-Based Learner. Technical report, Tilburg University, November 2003.

[54] T. M. Cover and P. E. Hart. Nearest Neighbor Pattern Classification. In *IEEE Trans. Inform. Theory*, pages 21–27, 1967.

[55] M. Rotaru. Dialog Act Tagging using Memory-Based Learning. Technical report, University of Pittsburgh, Spring 2002. Term Project in. Dialog Systems.

[56] D. W. Aha, D. Kibler, and M. K. Albert. Instance-Based Learning Algorithms. *Machine Learning*, 6(1):37–66, January 1991.

[57] P. Lendvai, A. van den Bosch, and Krahmer E. Machine Learning for Shallow Interpretation of User Utterances in Spoken Dialogue Systems. In *EACL-03 Workshop on Dialogue Systems: Interaction, Adaptation and Styles Management*, pages 69–78, Budapest, Hungary, 2003.

[58] E. Brill. *A Corpus-Based Approach to Language Learning*. PhD thesis, University of Pennsylvania, Philadelphia, USA, 1993.

[59] J. Woller. *The Basics of Monte Carlo Simulations*. University of Nebraska-Lincoln, Spring 1996. http://www.chem.unl.edu/zeng/joy/mclab/mcintro.html.

[60] K. Samuel, S. Carberry, and K. Vijay-Shanker. Dialogue Act Tagging with Transformation-Based Learning. In *17th international conference on Computational linguistics*, volume 2, pages 1150–1156, Montreal, Quebec, Canada, 10-14 August 1998. Association for Computational Linguistics, Morristown, NJ, USA.

[61] A. Van den Bosch, E. Krahmer, and M. Swerts. Detecting problematic turns in human-machine interactions: Rule-induction versus memory-based learning approaches. In *39th Meeting of the Association for Computational Linguistics*, pages 499–506, Toulouse, France, 2001.

[62] W. Cohen. Learning Trees and Rules with Set-valued Features. In *13th National Conference on Artificial Intelligence (AAAI-96)*, volume 1, pages 709–716, Portland, Oregon, 1996. AAAI Press.

[63] G. Tur, U. Guz, and D. Hakkani-Tür. Model adaptation for dialog act tagging. 2006.

[64] M. Zimmermann, D. Hakkani-Tür, E. Shriberg, and A. Stolcke. *Machine Learning for Multimodal interaction*, volume 4299 of *Lecture Notes in Computer Science*, chapter Text Based Dialog Act Classification for Multiparty Meetings, pages 190–199. Springer Berlin/Heidelberg, 2006.

Pavel Král is is graduated from University of West Bohemia at Dept. of Computer Science & Engineering in Plzeň (Czech Republic) and from Henri Poincaré University in Nancy (France) in 2007. He is now a lecturer, researcher at the University of West Bohemia. He is also a member of the Speech Group at LORIA-INRIA in Nancy. His research domain is on the speech recognition, more precisely on automatic dialogue act recognition. He received his M.Sc. degree in 1999 with honours in Dept. Informatics & Computer Science at the University of West Bohemia.

Christophe Cerisara is graduated from the engineering school ENSIMAG in computer science in Grenoble in 1996, and obtained the Ph.D. at the Institut National Polytechnique de Lorraine in 1999. He worked as a researcher from 1999 to 2000 at Panasonic Speech Technology Laboratory in Santa Barbara. He is now a research scientist at CNRS, and belongs to the Speech Group in LORIA. His research interests include multi-band models and robust automatic speech recognition to noise. He is the author or co-author of more than forty scientific publications.